March, 1956 BU-64-M

A FACTORIAL ANALYSIS OF VARIANCE MODEL AND ITS APPLICATION IN GENETICS

Douglas S. Robson
Cornell University

Introduction

Kempthorne (1954) has indicated that his construction of a model for genotypic values in a diploid population at equilibrium under random mating is based upon a model for the analysis of variance in factorial experiments. We shall here show the construction of the analysis of variance model and indicate explicitly its application to the genetic problem.

An Analysis of Variance Model

We suppose that individuals in a population are subjected to "treatments" consisting of combinations of levels of N factors, say factors A^1, A^2, ..., and A^N. The level of factor A^i assigned to an individual is determined by chance, independently for each i = 1, 2, ..., N, and may be either A^i_1, A^i_2, ..., or $A^i_{m_i}$. The probability that an individual will receive the particular level $A^i_{x_i}$ of factor A^i is written

$$(1) \qquad P\left\{A^i_{X_i} = A^i_{x_i}\right\} = p^i_{x_i}, \quad \sum_{x_i=1}^{m_i} p^i_{x_i} = 1.$$

We shall follow the conventional practice of using the upper case symbol X_i to designate the chance variable and the lower case x_i to denote a specific numerical value which may be taken on by the chance variable. Since we operate only on the subscripts it is useless to carry along the symbol $A^i_{x_i}$ and we shall write simply x_i, instead. Thus, we write (1) as

$$P\left\{X_i = x_i\right\} = p_{x_i}$$

for the probability that an individual will receive the level x_i of the i'th factor. By the assumption of independence, the probability that an individual

will receive the particular combination $x_1, x_2, \ldots, x_N$ of levels of the N factors is

$$(2) \qquad P\left\{X_1=x_1, X_2=x_2, \ldots, X_N=x_N\right\} = P\left\{X_1=x_1\right\} P\left\{X_2=x_2\right\} \ldots P\left\{X_N=x_N\right\}.$$

To further simplify the notation we shall denote the set $(x_1, x_2, \ldots, x_N)$ by the symbol x, so that (2) becomes

$$P\left\{X = x\right\} = \prod_{x_i \text{ in } x} P\left\{X_i = x_i\right\}.$$

We now assume the existence of a real- and single-valued function g(x), defined for all $x = (x_1, \ldots, x_N)$, and we shall refer to g(x) as the average "yield" under "treatment" x. The construction of an analysis of variance model for the chance variable g(X) then consists of defining the various "treatment" effects on "yield" in the population. The average effect of a given level x_i of the i'th factor is defined to be the difference between the population mean and the mean of all individuals carrying the level x_i. Denoting this effect by $f(x_i)$, we then have the definition

$$(3) \qquad f(x_i) = E\left\{g(X) \,\middle|\, X_i = x_i\right\} - E\left\{g(X)\right\}$$

where the operator E denotes expectation; thus, by (2),

$$f(x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N} g(x_1,\ldots,x_N) P\left\{X_1=x_1\right\} \ldots P\left\{X_{i-1}=x_{i-1}\right\} P\left\{X_{i+1}=x_{i+1}\right\} \ldots P\left\{X_N=x_N\right\}$$

$$- \sum_{x_1,\ldots,x_N} g(x_1,\ldots,x_N) P\left\{X_1= x_1\right\} \ldots P\left\{X_N= x_N\right\}.$$

Since $E\left\{g(X)\middle|X_i = x_i\right\}$ depends only on x_i we shall abbreviate this by writing $E\left\{g(X) \,\middle|\, X_i = x_i\right\} = \bar{g}(x_i)$ and, similarly, we define $E\left\{g(X)\right\} = \bar{\bar{g}}$, so that (3) becomes

$$(4) \qquad f(x_i) = \bar{g}(x_i) - \bar{\bar{g}}.$$

If the factors i and j did not interact in their effect upon yield; i.e., if the effect on yield of any given level x_i of factor i were the same regardless of the level of factor j present in the treatment then the average yield $\bar{g}(x_i,x_j)$ of all individuals receiving level x_i of factor i and x_j of factor j would be $\bar{g} + f(x_i) + f(x_j)$. The actual difference between $\bar{g}(x_i,x_j)$ and $\bar{g} + f(x_i) + f(x_j)$ is therefore defined to be the interaction effect of x_i and x_j and is denoted by $f(x_i,x_j)$; thus,

$$f(x_i,x_j) = E\left\{ g(X) \mid X_i = x_i,\ X_j = x_j \right\} - f(x_i) - f(x_j) - \bar{g}$$

$$= \bar{g}(x_i,x_j) - f(x_i) - f(x_j) - \bar{g} \tag{5}$$

or, by (4),

$$f(x_i,x_j) = \bar{g}(x_i,x_j) - \bar{g}(x_i) - \bar{g}(x_j) + \bar{g}. \tag{6}$$

If factors i and j interact in their effect upon yield but their interaction effect is the same regardless of the level of factor k present in the treatment then we would have, for every x_k,

$$E\left\{ g(X) \mid x_i,x_j,x_k \right\} - \left[E\left\{ g(X) \mid x_i,x_k \right\} - E\left\{ g(X) \mid x_k \right\} \right] - \left[E\left\{ g(X) \mid x_j,x_k \right\} - E\left\{ g(X) \mid x_k \right\} \right] - E\left\{ g(X) \mid x_k \right\}$$

$$= E\left\{ g(X) \mid x_i,x_j \right\} - \left[E\left\{ g(X) \mid x_i \right\} - E\left\{ g(X) \right\} \right] - \left[E\left\{ g(x) \mid x_j \right\} - E\left\{ g(X) \right\} \right] - E\left\{ g(X) \right\}.$$

In general, however, the interaction effect of x_i and x_j is not constant for all x_k and the above equality does not hold. We therefore define the interaction effect $f(x_i, x_j, x_k)$ to be the difference between the left and right sides above; thus,

$$f(x_i,x_j,x_k) = \Big(E\left\{ g(X) \mid x_i,x_j,x_k \right\} - \left[E\left\{ g(X) \mid x_i,x_k \right\} - E\left\{ g(X) \mid x_k \right\} \right] - \left[E\left\{ g(X) \mid x_j,x_k \right\} - E\left\{ g(x) \mid x_k \right\} \right] - E\left\{ g(X) \mid x_k \right\} \Big)$$

$$- \Big(E\left\{ g(x) \middle| x_i, x_j \right\} - [\, E\left\{ g(x) \middle| x_i \right\} - E\left\{ g(X) \right\}]$$
$$- [\, E\left\{ g(X) \middle| x_j \right\} - E\left\{ g(X) \right\}] - E\left\{ g(X) \right\} \Big)$$
$$= [\, \bar{g}(x_i, x_j, x_k) - \left\{ \bar{g}(x_i, x_k) - \bar{g}(x_k) \right\} - \left\{ \bar{g}(x_j, x_k) - \bar{g}(x_k) \right\} - \bar{g}(x_k)]$$
$$- [\, \bar{g}(x_i, x_j) - \left\{ \bar{g}(x_i) - \bar{g} \right\} - \left\{ \bar{g}(x_j) - \bar{g} \right\} - \bar{g} \,]$$

or

$$(7) \quad f(x_i, x_j, x_k) = \bar{g}(x_i, x_j, x_k) - \bar{g}(x_i, x_j) - \bar{g}(x_i, x_k) - \bar{g}(x_j, x_k) + \bar{g}(x_i) + \bar{g}(x_j) + \bar{g}(x_k) - \bar{g}$$

or, since $\bar{g}(x_i, x_j) = f(x_i, x_j) + f(x_i) + f(x_j) + \bar{g}$, by (5), and $\bar{g}(x_i) = f(x_i) + \bar{g}$, by (4), we may write

$$(8) \quad f(x_i, x_j, x_k) = \bar{g}(x_i, x_j, x_k) - f(x_i, x_j) - f(x_i, x_k) - f(x_j, x_k) - f(x_i) - f(x_j) - f(x_k) - \bar{g}$$

Extending this reasoning, we define the effect $f(x_{i_1}, \ldots, x_{i_v})$ as

$$(9) \quad f(x_{i_1}, \ldots, x_{i_v}) = \sum_{t=0}^{v} (-1)^{v-t} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } (x_{i_1}, \ldots, x_{i_v})} \bar{g}(x_{h_1}, \ldots, x_{h_t})$$

with the understanding that for $t = 0$ the set $(x_{h_1}, \ldots, x_{h_t})$ is the (unique) empty set and $\bar{g}(x_{h_1}, \ldots, x_{h_t}) \big|_{t=0} = \bar{g}$. To show that (9) may also be written as

$$(10) \quad f(x_{i_1}, \ldots, x_{i_v}) = \bar{g}(x_{i_1}, \ldots, x_{i_v}) - \sum_{t=0}^{v-1} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } (x_{i_1}, \ldots, x_{i_v})} f(x_{h_1}, \ldots, x_{h_t})$$

again with the convention that $f(x_{h_1}, \ldots, x_{h_s})\Big|_{s=0} = \bar{g}$, we observe from (9) that

$$f(x_{h_1}, \ldots, x_{h_t}) = \sum_{s=0}^{t} (-1)^{t-s} \sum_{(x_{k_1}, \ldots, x_{k_s}) \text{ in } (x_{h_1}, \ldots, x_{h_t})} \bar{g}(x_{k_1}, \ldots, x_{k_s})$$

so that

$$\sum_{t=0}^{\upsilon-1} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } (x_{i_1}, \ldots, x_{i_\upsilon})} f(x_{h_1}, \ldots, x_{h_t})$$

$$= \sum_{t=0}^{\upsilon-1} \sum_{(x_{h_1},\ldots,x_{h_t}) \text{ in } (x_{i_1},\ldots,x_{i_\upsilon})} \left\{ \sum_{s=0}^{t} (-1)^{t-s} \sum_{(x_{k_1},\ldots,x_{k_s}) \text{ in } (x_{h_1},\ldots,x_{h_t})} \bar{g}(x_{k_1},\ldots,x_{k_s}) \right\}.$$

But every set $(x_{k_1}, \ldots, x_{k_s})$ is a subset of $\binom{\upsilon-s}{t-s}$ different subsets of t elements from the set $(x_{i_1}, \ldots, x_{i_\upsilon})$, hence the right side above may be written

$$= \sum_{s=0}^{\upsilon-1} \sum_{(x_{k_1}, \ldots, x_{k_s}) \text{ in } (x_{i_1}, \ldots, x_{i_\upsilon})} \bar{g}(x_{k_1}, \ldots, x_{k_s}) \left\{ \sum_{t=s}^{\upsilon-1} (-1)^{t-s} \binom{\upsilon-s}{t-s} \right\}$$

and since

$$\sum_{t=s}^{\upsilon-1} (-1)^{t-s} \binom{\upsilon-s}{t-s} = \sum_{t=s}^{\upsilon} (-1)^{t-s} \binom{\upsilon-s}{t-s} - (-1)^{\upsilon-s} \binom{\upsilon-s}{\upsilon-s} = -(-1)^{\upsilon-s}$$

then (9) and (10) are equivalent.

The analysis of variance model is obtained from the above results by taking $\upsilon = N$ in equation (10). This gives

$$f(x_1, \ldots, x_N) = \bar{g}(x_1, \ldots, x_N) - \sum_{t=0}^{N-1} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } (x_1, \ldots, x_N)} f(x_{h_1}, \ldots, x_{h_t})$$

or, since $(x_1, \ldots, x_N) = x$ and $\bar{g}(x) = g(x)$,

$$g(x) = \sum_{t=0}^{N} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } x} f(x_{h_1}, \ldots, x_{h_t})$$

Hence, the chance variable $g(X)$ may be expressed in terms of the linear model

$$g(X) = \sum_{t=0}^{N} \sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, N)} f(X_{h_1}, \ldots, X_{h_t}) . \tag{11}$$

Since all effects $f(X_{h_1}, \ldots, X_{h_t})$ are random variables for $t > 0$ then, by definition, in order to demonstrate that (11) is an analysis of variance model we must show that all effects other than $\bar{g}$ have zero mean and are uncorrelated. To accomplish this we use the relation (9); thus, the mean value of $f(X_{i_1}, \ldots, X_{i_\upsilon})$ is

$$E\left\{f(X_{i_1}, \ldots, X_{i_\upsilon})\right\} = \sum_{t=0}^{\upsilon} (-1)^{\upsilon-t} \sum_{(h_1, \ldots, h_t) \text{ in } (i_1, \ldots, i_\upsilon)} E\left\{\bar{g}(X_{h_1}, \ldots, X_{h_t})\right\} .$$

Since $E\left\{\bar{g}(X_{h_1}, \ldots, X_{h_t})\right\} = \bar{g}$ identically and since there are $\binom{\upsilon}{t}$ subsets of $(i_1, \ldots, i_\upsilon)$ which contain t elements, then the right side above becomes

$$= \bar{g} \sum_{t=0}^{\upsilon} (-1)^{\upsilon-t} \binom{\upsilon}{t}$$

$$= \begin{cases} 0 \text{ for } \upsilon > 0 \\ \bar{g} \text{ for } \upsilon = 0 \end{cases}$$

To show that for $(i_1, \ldots, i_s) \neq (j_1, \ldots, j_t)$, $f(X_{i_1}, \ldots, X_{i_s})$ and $f(X_{j_1}, \ldots, X_{j_s})$ are uncorrelated it is then only necessary to show that

$$E\left\{f(X_{i_1}, \ldots, X_{i_s}) \cdot f(X_{j_1}, \ldots, X_{j_t})\right\} = 0.$$

This we accomplish by first noting that

$$(12) \qquad E\left\{f(X_{i_1}, \ldots, X_{i_s})f(X_{j_1}, \ldots, X_{j_s})\right\}$$

$$= E\left[f(X_{i_1}, \ldots, X_{i_s}) \cdot E\left\{f(X_{j_1}, \ldots, X_{j_t}) \,\middle|\, X_{i_1}, \ldots, X_{i_s}\right\}\right]$$

$$= E\left[f(X_{i_1}, \ldots, X_{i_s}) \sum_{q=0}^{t} (-1)^{t-q} \sum_{(h_1, \ldots, h_q) \text{ in } (j_1, \ldots, j_t)} E\left\{\bar{g}(X_{h_1}, \ldots, X_{h_q}) \,\middle|\, X_{i_1}, \ldots, X_{i_s}\right\}\right].$$

We next observe that if $(X_{h_1}, \ldots, X_{h_q}) \cap (X_{i_1}, \ldots, X_{i_s})$ denotes the (intersection) chance variables which occur in both sets $(X_{i_1}, \ldots, X_{i_s})$ and $(X_{h_1}, \ldots, X_{h_s})$ then, since

$$E\left\{\bar{g}(X_{h_1}, \ldots, X_{h_q}) \,\middle|\, X_{i_1}, \ldots, X_{i_s}\right\} = E\left[E\left\{g(X) \,\middle|\, X_{h_1}, \ldots, X_{h_q}\right\} \,\middle|\, X_{i_1}, \ldots, X_{i_s}\right]$$

$$= E\left\{g(X) \,\middle|\, (X_{h_1}, \ldots, X_{h_q}) \cap (X_{i_1}, \ldots, X_{i_s})\right\},$$

the right side of (12) becomes

$$= E\left[f(X_{i_1}, \ldots, X_{i_s}) \sum_{q=0}^{t} (-1)^{t-q} \sum_{(h_1, \ldots, h_q) \text{ in } (j_1, \ldots, j_t)} E\left\{g(X) \,\middle|\, (X_{h_1}, \ldots, X_{h_q}) \cap (X_{i_1}, \ldots, X_{i_s})\right\}\right].$$

Now let $(v_1, \ldots, v_n) = (i_1, \ldots, i_s) \cap (j_1, \ldots, j_t)$ and consider an arbitrary subset $(v_{k_1}, \ldots, v_{k_m})$ of the set $(v_1, \ldots, v_n)$. Then among the $\binom{t}{q}$ subsets of $(j_1, \ldots, j_t)$ which contain q elements there are $\binom{t-n}{q-m}$ which contain the set $(v_{k_1}, \ldots, v_{k_m})$ and no other elements of the set $(v_1, \ldots, v_n)$, $m \leq q \leq m + t - n$. Hence, the coefficient of $E\left\{g(X) \,\middle|\, X_{v_{k_1}}, \ldots, X_{v_{k_m}}\right\}$ in the above square bracket is

$$(13)\qquad f(X_{i_1}, \ldots, X_{i_s}) \sum_{q=m}^{m+t-n} (-1)^{t-q} \binom{t-n}{q-m}$$

$$= (-1)^{n-m} f(X_{i_1}, \ldots, X_{i_s}) \sum_{q=m}^{m+t-n} (-1)^{(t-n)-(q-m)} \binom{t-n}{q-m}$$

which vanishes when $t > n$; i.e., when the set $(j_1, \ldots, j_t)$ is not a subset of $(i_1, \ldots, i_s)$. However, since $(i_1, \ldots, i_s) \neq (j_1, \ldots, j_t)$ then if $(j_1, \ldots, j_t)$ is a subset of $(i_1, \ldots, i_s)$ it is a proper subset (i.e., not equal to) and the above argument may be applied replacing (12) by

$$E\left\{f(X_{i_1}, \ldots, X_{i_s}) f(X_{j_1}, \ldots, X_{j_t})\right\}$$

$$= E\left[f(X_{j_1}, \ldots, X_{j_t}) E\left\{f(X_{i_1}, \ldots, X_{i_s}) \,\middle|\, X_{j_1}, \ldots, X_{j_t}\right\}\right].$$

We may assume, therefore, that $t > n$. Since $(v_{k_1}, \ldots, v_{k_m})$ was an arbitrary subset of $(v_1, \ldots, v_n)$ then (13) holds for all subsets of $(v_1, \ldots, v_n)$ so the right side of (12) vanishes completely giving the desired result

$$E\left\{f(X_{i_1}, \ldots, X_{i_s}) f(X_{j_1}, \ldots, X_{j_t})\right\} = 0 \text{ when } (i_1, \ldots, i_s) \neq (j_1, \ldots, j_t).$$

The variance σ_g^2 of the chance variable $g(X)$ is therefore

$$(14)\qquad \sigma_g^2 = \sum_{t=1}^{N} \sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, N)} E\left\{f(X_{h_1}, \ldots, X_{h_t})\right\}^2.$$

Application to the Random Mating Problem

A special case of the above analysis of variance model is obtained by imposing additional restrictions upon the underlying probability model. We have assumed that the chance variables $X_1, \ldots, X_N$ are independently distributed; the additional restrictions which characterize the random mating

problem are, first, that N = rn, where r and n are integers, and second, for every i, $1 \leq i \leq n$, the chance variables X_i, X_{n+i}, ..., $X_{(r-1)n+i}$ are identically distributed. In this special case, then, the N factors consist of r identical sets of n factors. In the genetic problem the n factors represent n loci and the m_i levels, A_1^i, ..., $A_{m_i}^i$, of the i'th factor represent the different alleles present at this locus in the population. The integer r, then, represents the number of repetitions of the haploid chromosomal complement present in the organisms; for $r > 2$ the organisms must be autopolyploids. We shall consider only the case, r = 2, of a diploid population at equilibrium under random mating.

Let $A_{x_1}^1$, ..., $A_{x_n}^n$ denote the set of genes contributed by the sire and $A_{x_{n+1}}^1$, ..., $A_{x_{2n}}^n$ denote the set of genes contributed by the dam, so that an individual's genotype is written $A_{x_1}^1$, $A_{x_{n+1}}^1$, ..., $A_{x_n}^n$ $A_{x_{2n}}^n$. The genotypic value g(x) of the genotype defined by $x = (x_1, \ldots, x_{2n})$ may then be expressed as

$$g(x) = \sum_{t=0}^{2n} \sum_{(x_{h_1}, \ldots, x_{h_t}) \text{ in } x} f(x_{h_1}, \ldots, x_{h_t})$$

and the chance variable g(X) as

$$g(X) = \sum_{t=0}^{2n} \sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, 2n)} f(X_{h_1}, \ldots, X_{h_t})$$

with the understanding that $f(x_{h_1}, \ldots, x_{h_t})\big|_{t=0} = f(X_{h_1}, \ldots, X_{h_t})\big|_{t=0} = \bar{g}$. Since X_i and X_{n+i} are identically distributed then for each t, $0 \leq t \leq 2n$, certain of the chance variables $f(X_{h_1}, \ldots, X_{h_t})$ are also identically distributed. For example, for t = 2 the 4 chance variables $f(X_i, X_j)$, $f(X_i, X_{n+j})$,

$f(X_{n+i}, X_j)$ and $f(X_{n+i}, X_{n+j})$, $1 \leq i < j \leq n$, are identically distributed, and therefore have the same variance; there are $\binom{n}{2}$ such sets of 4 identically distributed chance variables among the $\binom{2n}{2}$ chance variables in the sum

$$\sum_{(h_1, h_2) \text{ in } (1, \ldots, 2n)} f(X_{h_1}, X_{h_2})$$

and the remaining $n = \binom{2n}{2} - 4\binom{n}{2}$ chance variables are of the form $f(X_i, X_{n+i})$, $1 \leq i \leq n$. We may therefore write

$$\sum_{(h_1, h_2) \text{ in } (1, \ldots, 2n)} E\left\{f(X_{h_1}, X_{h_2})\right\}^2 = 4 \sum_{(i,j) \text{ in } (1, \ldots, n)} E\left\{f(X_i, X_j)\right\}^2$$

$$+ \sum_{(i) \text{ in } (1, \ldots, n)} E\left\{f(X_i, X_{n+i})\right\}^2.$$

The chance effect $f(X_i, X_j)$, $1 \leq i < j \leq n$, represents an interaction effect between two genes at different loci as compared to $f(X_i, X_{n+i})$ which represents an interaction between two alleles at the same locus. Because of their genetic significance these two types of effects have been given different names, $f(X_i, X_j)$ being called an "additive x additive" effect and $f(X_i, X_{n+i})$ a "dominance" effect. The expression

$$4 \sum_{(i,j) \text{ in } (1, \ldots, n)} E\left\{f(X_i, X_j)\right\}^2$$

is then called $\sigma^2_{A^2}$, the "additive x additive" component of genetic variance, and,

$$\sum_{(i) \text{ in } (1, \ldots, n)} E\left\{f(X_i, X_{n+i})\right\}^2$$

is called σ^2_D, the "dominance" component of variance. In general, it can be seen that

$$\sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, 2n)} f(X_{h_1}, \ldots, X_{h_t})$$

contains $2^{t-2\upsilon}$ chance variables which have the same distribution as

$f(X_{i_1}, \ldots, X_{i_\upsilon}, X_{i_{\upsilon+1}}, \ldots, X_{i_{t-\upsilon}}, X_{n+i_1}, \ldots, X_{n+i_\upsilon})$, $1 \leq i_1 < \ldots < i_{t-\upsilon} \leq n$.

Such an effect is called an "additive x ... x additive x dominance x ... x dominance" effect, where the word "additive" appears $t - 2\upsilon$ times and "dominance" appears υ times. Since

$$\sum_{(h_1,\ldots,h_t) \text{ in } (1,\ldots,2n)} E\left\{f(X_{h_1},\ldots,X_{h_t})\right\}^2$$

$$= \sum_{\upsilon=\max(0,t-n)}^{[\frac{t}{2}]} 2^{t-2\upsilon} \sum_{(i_1,\ldots,i_{t-\upsilon}) \text{ in } (1,\ldots,n)} E\left\{f(X_{i_1},\ldots,X_{i_\upsilon},\ldots,X_{i_{t-\upsilon}},X_{n+i_1},\ldots,X_{n+i_\upsilon})\right\}^2$$

where $[\frac{t}{2}]$ = the largest integer in $\frac{t}{2}$, then the expression

$$2^{t-2\upsilon} \sum_{(i_1,\ldots,i_{t-\upsilon}) \text{ in } (1,\ldots,n)} E\left\{f(X_{i_1},\ldots,X_{i_\upsilon},\ldots,X_{i_{t-\upsilon}},X_{n+i_1},\ldots,X_{n+i_\upsilon})\right\}^2$$

is called $\sigma^2_{A^{t-2\upsilon}D^\upsilon}$. Thus, the genetic variance σ^2_g because

$$\sigma^2_g = \sum_{t=1}^{2n} \sum_{\upsilon=\max(0,t-n)}^{[\frac{t}{2}]} \sigma^2_{A^{t-2\upsilon}D^\upsilon}$$

or

$$\sigma^2_g = \sum_{\substack{0 \leq r,\, s \leq n \\ 1 \leq r + 2s \leq 2n}} \sigma^2_{A^r D^s} .$$

References

Kempthorne, O. (1954). The correlation between relatives in a random mating population. Proc. Roy. Soc. London, B, 143:103-113.

PROBLEM FOR STUDENTS OF PLANT BREEDING 214: SIMPLE AUTOPOLYPLOIDY

If in the analysis of variance model N = rn and X_i, X_{i+n}, ..., $X_{i+(r-1)n}$ are r identically distributed chance variables, i=1, 2, ..., n, then the model may be regarded as describing genotypic values in a simple autopolyploid population which is at equilibrium under random mating. The expression for genetic variance,

$$(1) \qquad \sigma_g^2 = \sum_{t=1}^{rn} \sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, rn)} E\left\{ f(X_{h_1}, \ldots, X_{h_t}) \right\}^2$$

may again be reduced, as in the diploid case r = 2, in that for each t certain of the chance variables $f(X_{h_1}, \ldots, X_{h_t})$ are identically distributed and hence have the same variance. Thus, among the $\binom{rn}{t}$ chance variables in the sum

$$\sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, rn)} f(X_{h_1}, \ldots, X_{h_t})$$

there are $\prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha}$ chance variables which have the same distribution as

$$f(X_{h_{11}}, \ldots, X_{h_{1k_1}}, X_{h_{21}}, X_{h_{21}+n}, \ldots, X_{h_{2k_2}}, X_{h_{2k_2}+n}, \ldots$$

$$\ldots, X_{h_{r1}}, \ldots X_{h_{r1}+(r-1)n}, \ldots \underset{r}{X_{h_{rk_r}}}, \ldots X_{h_{rk_r}+(r-1)n})$$

where $1 \leq h_{\alpha 1} < \ldots < h_{\alpha k_\alpha} \leq n$ for $\alpha = 1, 2, \ldots, r$ and $\sum_{\alpha=1}^{r} \alpha k_\alpha = t$. We therefore define the genetic variance component

$$\sigma^2_{k_1, \ldots, k_r} = \prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha} \sum_{(h_{11}, \ldots, n_{rk_r}) \text{ in } (1, \ldots, n)} E\Big\{ f(X_{h_{11}}, \ldots, X_{h_{1k_1}}, X_{h_{21}}, X_{h_{21}+n}, \ldots$$

$$X_{h_{2k_2}}, X_{h_{2k_2}+n}, \ldots\ldots, X_{h_{r1}}, \ldots, X_{h_{r1}+(r-1)n}, \ldots, X_{h_{rk_r}}, \ldots, X_{h_{rk_r}+(r-1)n}) \Big\}^2 .$$

The problem is to prove that

$$\sigma_g^2 = \sum_{\substack{0 \le k_1, \ldots, k_r \le n \\ 1 \le \sum_{\alpha=1}^{r} \alpha k_\alpha \le rn}} \sigma^2_{k_1, \ldots, k_r}$$

Clearly, by (1), it will suffice to show that for arbitrary t, $1 \le t \le rn$,

$$(2) \qquad \sum_{(h_1, \ldots, h_t) \text{ in } (1, \ldots, rn)} E\left\{f(X_{h_1}, \ldots, X_{h_t})\right\}^2 = \sum_{\substack{0 \le k_1, \ldots, k_t \le n \\ \sum_{\alpha=1}^{r} \alpha k_\alpha = t}} \sigma^2_{k_1, \ldots, k_r}$$

or, in other words, that because of our definition of $\sigma^2_{k_1, \ldots, k_r}$ the right sum of (2) accounts for each of the $\binom{rn}{t}$ terms appearing in the left sum of (2). The problem thus reduces to one of combinatorial analysis.

Motivation for this problem may be found in the paper "The correlation between relatives in a simple autotetraploid population" by O. Kempthorne, Genetics 40:168-174, 1955. Kempthorne treats the case r = 4, n = 1.

SOLUTION TO THE AUTOPOLYPLOIDY PROBLEM:

The set $(h_{11}, \ldots, h_{rk_r})$ contains $\sum_{\alpha=1}^{r} k_\alpha$ integers from the set $(1, \ldots, n)$. The number of different subsets of $\sum_{\alpha=1}^{r} k_\alpha$ integers which may be formed from the set $(1, \ldots, n)$ is $\binom{n}{\Sigma k_\alpha}$. Associated with each such subset of Σk_α integers is the collection of chance variables f which can be formed by partitioning the Σk_α subscripts $(h_{11}, \ldots, h_{rk_r})$ into r subsets containing $k_1, k_2, \ldots, k_r$ elements, respectively. The number of such partitions is

$$\frac{(\Sigma k_\alpha)!}{\prod (k_\alpha!)}$$

The variance component $\sigma^2_{k_1, \ldots, k_r}$ therefore accounts for

$$\prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha} \binom{n}{\Sigma k_\alpha} \frac{(\Sigma k_\alpha)!}{\prod (k_\alpha!)} = \prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha} \frac{n!}{\prod (k_\alpha!)(n-\Sigma k_\alpha)!}$$

of the $\binom{rn}{t}$ terms in the left sum of (2). It remains, then, only to show that

$$\sum_{\substack{0 \le k_1, \ldots, k_r \le n \\ \Sigma \alpha k_\alpha = t}} \prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha} \frac{n!}{\prod (k_\alpha!)(n-\Sigma k_\alpha)!} = \binom{rn}{t}$$

This we accomplish by comparing coefficients of b^t on both sides of

$$(1+b)^{rn} = [(1+b)^r]^{\Sigma k_\alpha + n - \Sigma k_\alpha}$$

On the left side we get $\binom{rn}{t}$ as the coefficient of b^t. The right side we expand as

$$[(1+b)^r]^{\Sigma k_\alpha + n - \Sigma k_\alpha} = \left[1 + \binom{r}{1}b + \binom{r}{2}b^2 + \ldots + \binom{r}{r}b^r\right]^{\Sigma k_\alpha + n - \Sigma k_\alpha}$$

$$= \sum_{0 \le k_1, \ldots, k_r \le n} \frac{n!}{k_1! \ldots k_r! (n-\Sigma k_\alpha)!} \prod \binom{r}{\alpha}^{k_\alpha} b^{\Sigma \alpha k_\alpha}$$

Hence the coefficient of $b^{\Sigma \alpha k_\alpha}\Big|_{\Sigma \alpha k_\alpha = t}$ on the right side is

$$\sum_{\substack{0 \le k_1, \ldots, k_r \le n \\ \Sigma \alpha k_\alpha = t}} \frac{n!}{k_1! \ldots k_r! (n-\Sigma k_\alpha)!} \prod_{\alpha=1}^{r} \binom{r}{\alpha}^{k_\alpha}$$

and the problem is solved.